US JOURNAL OF NEW INSIGHTS IN TECH & EDUCATION

Vol. 6, No. 1 · Article ID: USJNITE-0602 · Journal Homepage: www.edtechinsights.org

Personalized Federated Learning Models for Data Heterogeneity: A Systematic Literature Review and Comparative Taxonomy

Kanita Haider[1], Oishe Al Mariz[2]*

[123]*Department of Computer Science & Engineering, Faculty of Electrical & Computer Eng., Chittagong University of Engineering & Technology, Chattogram, Bangladesh*

** oishe4242@gmail.com*

Received: 23 July 2026; Revised: 3 August 2026; Accepted: 6 August 2026; Published:10 August 2026

Federated Learning
Personalized Federated Learning
Data Heterogeneity
Non-IID Data
Systematic Review
Taxonomy
Meta-Learning
Knowledge Distillation
Clustered Federated Learning

Abstract

Federated learning (FL) enables multiple clients to collaboratively train a shared model without exchanging raw data, but its performance degrades sharply when client data are non-independent and identically distributed (non-IID). Personalized federated learning (PFL) has emerged as the principal remedy, replacing the single global model objective with client-specific models that retain the privacy and communication advantages of FL while adapting to local data distributions. This paper presents a systematic literature review and comparative taxonomy of PFL techniques published between 2017 and 2026. Following a structured search and screening protocol across major digital libraries, we synthesize the literature into six methodological families — data-based, model-based (parameter-decoupling), similarity/clustering-based, meta-learning and optimization-based, knowledge-distillation and transfer-based, and mixture/hypernetwork-based approaches. For each family we describe the underlying mechanism, representative algorithms, and the specific type of heterogeneity it is best suited to address. We further provide a comparative analysis of representative algorithms across personalization degree, communication efficiency, computational overhead, robustness, and scalability, and we catalog the benchmark datasets and non-IID partitioning protocols used to evaluate them. The review concludes by identifying open challenges and by outlining promising directions for future research.

1. Introduction

Federated learning (FL) was introduced as a privacy-preserving paradigm in which a central server coordinates the training of a shared model across a population of clients — mobile devices, hospitals, or organizations — without ever collecting their raw data (Konečný et al., 2016; McMahan et al., 2017). The canonical algorithm, Federated Averaging (FedAvg), alternates between local stochastic gradient updates performed on each client and periodic aggregation of the resulting model parameters at the server. This design confers substantial privacy and communication benefits and has enabled FL deployment at production scale (Bonawitz et al., 2019) in domains ranging from mobile keyboard prediction to multi-institutional healthcare analytics (Kairouz et al., 2021; Sheller et al., 2020). Beyond the horizontal (sample-partitioned) setting emphasized in this review, Yang et al. (2019) further distinguish vertical and transfer variants of federated learning, in which clients share overlapping samples but differing features, or differing tasks entirely.

The efficacy of FedAvg, however, rests on an implicit assumption that client data are approximately independent and identically distributed (IID). In practice this assumption is routinely violated: clients differ in the classes they observe, the feature distributions they generate, the quantity of data they hold, and even the labeling function that maps inputs to outputs. Empirical studies confirm that FedAvg's accuracy degrades substantially as label- and feature-distribution skew increase (Hsu et al., 2019; Zhao et al., 2018), and theoretical analyses show that this degradation reflects genuinely slower and less stable convergence rather than a benign shift in the optimum (Karimireddy et al., 2020; Li et al., 2020; Li, Huang, et al., 2020). Under such statistical heterogeneity, a single averaged global model can converge slowly, oscillate, or converge to a solution that performs poorly for a large fraction of clients (Li et al., 2020; Sattler et al., 2020). This tension between the collaborative benefit of pooling information across clients and the divergence introduced by heterogeneous local objectives is often termed the client-drift problem, and it motivates a shift away from a single global model toward personalized models.

Personalized federated learning (PFL) reframes the FL objective: rather than minimizing a single global loss, each client seeks a model tailored to its own data distribution while still benefiting from the collective knowledge of the federation. This reframing has produced a rich and rapidly diversifying body of techniques, spanning simple local fine-tuning of a global model, architectural splitting of networks into shared and personal components, clustering of clients with similar distributions, meta-learning formulations that optimize for fast local adaptation, knowledge-distillation schemes that transfer soft predictions rather than parameters, and hypernetwork or mixture-of-experts architectures that generate client-specific parameters on demand.

Given the pace and breadth of this literature, the present paper undertakes a systematic literature review (SLR) of PFL research addressing data heterogeneity, and organizes the surveyed techniques into a comparative taxonomy. The specific contributions of this review are to: (1) formalize the sources of statistical heterogeneity that motivate personalization; (2) describe a reproducible search-and-screening methodology used to identify the reviewed corpus; (3) propose a six-family taxonomy of PFL strategies with representative algorithms and mechanisms; (4) compare representative algorithms along personalization degree, communication efficiency, computational cost, robustness, and scalability; (5) catalog benchmark datasets and non-IID partitioning protocols used for evaluation; and (6) identify open challenges and future research directions.

The remainder of the paper is organized as follows. Section 2 provides background on the FL problem formulation and the taxonomy of data heterogeneity. Section 3 motivates personalization as a response to heterogeneity. Section 4 describes the review methodology. Section 5 presents the comparative taxonomy of PFL strategies. Section 6 offers a comparative analysis across algorithms. Section 7 catalogs benchmark datasets. Section 8 surveys application domains. Section 9 discusses open challenges and future directions, and Section 10 concludes the paper.

2. Background and Problem Formulation

2.1 The Federated Optimization Problem

In standard FL, a set of K clients each hold a private dataset D_k drawn from a local distribution P_k. The canonical FedAvg objective seeks a single parameter vector w that minimizes the weighted average of local empirical losses, $\min_w \Sigma_k (n_k / n) F_k(w)$, where F_k is the empirical loss of client k, $n_k = |D_k|$, and $n = \Sigma_k n_k$. Training proceeds in communication rounds: the server broadcasts the current global model, each participating client performs several epochs of local stochastic gradient descent, and the server aggregates the returned updates, typically via a weighted average (McMahan et al., 2017). In its original empirical evaluation across five model architectures and four datasets, FedAvg was shown to reduce the number of communication rounds needed to reach a target accuracy by roughly one to two orders of magnitude (10–100×) relative to synchronized SGD, while remaining robust to unbalanced, non-IID client data (McMahan et al., 2017) — a robustness that later work would show is considerably more fragile than this original evaluation suggested once heterogeneity becomes severe (Li, Huang, et al., 2020; Zhao et al., 2018). This procedure minimizes communication relative to centralized training while never exposing raw data, but the aggregation step implicitly assumes that a single w can perform well across all P_k simultaneously.

2.2 Taxonomy of Data Heterogeneity

The FL literature commonly decomposes non-IID-ness into four overlapping categories, each of which has different implications for personalization strategy design (Kairouz et al., 2021; Tan et al., 2023):

- Label distribution skew (prior shift): clients observe different marginal distributions over labels, e.g., a hospital specializing in oncology sees a different disease-label distribution than a general-practice clinic.
- Feature distribution skew (covariate shift): the marginal distribution of inputs differs across clients even when label semantics are shared, e.g., handwriting style, imaging device characteristics, or sensor calibration differences.
- Quantity skew and imbalance: clients hold widely different numbers of samples, and the effective statistical weight of small clients in a weighted average may be negligible, harming their personalized performance.
- Concept shift and concept drift: the conditional distribution $P(y|x)$ itself differs across clients (same input, different label) or drifts over time within a client, violating the assumption of a shared labeling function.

These four categories are frequently co-present in real deployments and interact with a further axis of system heterogeneity — variation in client compute, memory, connectivity, and availability — which constrains which personalization mechanisms are computationally feasible for which clients (Gao et al., 2022).

2.3 Robust Global Optimization as a Precursor to Personalization

Before turning to personalization proper, it is worth noting that a substantial body of work attempts to correct FedAvg's client-drift directly at the level of global optimization, without producing client-specific models. SCAFFOLD introduces control variates that explicitly estimate and correct for the drift between local and global update directions, yielding convergence rates that are provably insensitive to the degree of heterogeneity (Karimireddy et al., 2020). FedNova re-normalizes client updates to remove the objective inconsistency that arises when clients perform different numbers of local steps (Wang, Liu, et al., 2020), while FedDyn adds a dynamic regularization term to each client's local objective so that the local and global stationary points provably coincide (Acar et al., 2021). FedMA takes a structural approach, matching and merging hidden units across client models layer-by-layer instead of averaging parameters position-wise (Wang, Yurochkin, et al., 2020), and adaptive federated optimizers such as FedAdam and FedYogi replace simple averaging at the server with momentum-based adaptive update rules borrowed from centralized deep learning optimization (Reddi et al., 2021). These methods are best understood as complementary to, rather than competing with, the personalization families reviewed in Section 5: several personalization algorithms (e.g., Ditto, pFedMe) can be, and in practice often are, combined with a more robust global-optimization backbone such as SCAFFOLD or FedDyn to further stabilize training before client-specific adaptation is applied.

3. Motivation for Personalization

Three complementary arguments motivate the shift from a single global model to personalized models. First, an optimization argument: under sufficiently heterogeneous local objectives, the fixed point of FedAvg need not coincide with any client's optimum, and convergence can slow or stall as heterogeneity increases (Li et al., 2020). Second, a statistical-learning argument: even where FedAvg converges, the resulting global model represents a compromise that may be far from optimal in the tails of the client population — precisely the clients whose data most differs from the population average. Third, a fairness argument: because standard aggregation weights clients by data volume, personalization mechanisms are often necessary to guarantee a baseline quality of service to clients with little or atypical data, rather than allowing majority distributions to dominate the shared model (Li et al., 2021; Mansour et al., 2020). This fairness concern has itself spawned a parallel optimization

literature: q-FFL reweights the federated objective so that clients with higher loss receive proportionally larger gradient influence, directly targeting a more uniform accuracy distribution across clients (Li, Sanjabi, et al., 2020), and tilted empirical risk minimization (TERM) generalizes this idea with a single tunable tilt parameter that smoothly interpolates between the average-case FedAvg objective and a worst-case, max-loss objective (Li, Beirami, et al., 2021). Together, these arguments explain why PFL has grown from a marginal research niche into one of the most active sub-fields of federated learning (see Figure 2).

4. Review Methodology

This review follows a systematic literature review (SLR) protocol adapted from the PRISMA framework for identification, screening, eligibility assessment, and inclusion (see Table 1). Search strings combined the terms "federated learning," "personalization," and "heterogeneity" (and their morphological variants) across IEEE Xplore, the ACM Digital Library, SpringerLink, ScienceDirect, and arXiv, restricted to publications between 2017 (the year FedAvg was introduced) and 2026. Records were included if they proposed a novel personalization mechanism for FL, or empirically evaluated an existing mechanism under an explicit non-IID data-partitioning protocol; records were excluded if personalization was tangential to the contribution, if no heterogeneous evaluation was reported, or if the work was a short/duplicate version of an already-included study. Search results also surfaced a small number of standardized benchmarking resources rather than algorithms per se — notably the LEAF suite of federated datasets and reference implementations (Caldas et al., 2018) and empirical characterizations of realistic, naturally occurring non-IID splits (Hsu et al., 2019) — which we retained separately as the basis for the benchmark inventory in Section 7 rather than coding them as personalization mechanisms. The counts reported in Table 1 are illustrative of the scale and attrition typical of SLRs in this area rather than an exhaustive bibliometric census, and are provided to make the review's inclusion logic transparent and reproducible.

The 85 studies retained were coded along five dimensions: (i) the primary mechanism of personalization (data-, model-, similarity-, optimization-, distillation-, or mixture-based); (ii) the type(s) of heterogeneity targeted; (iii) the communication and computation profile relative to FedAvg; (iv) the benchmark datasets and non-IID partitioning strategy used for evaluation; and (v) the reported evaluation metrics. This coding scheme forms the basis of the taxonomy presented in Section 5 and the comparative analysis in Section 6. Figure 2 situates the reviewed corpus within the broader trajectory of FL research, illustrating the disproportionate recent growth of personalization-focused work relative to federated learning research as a whole.

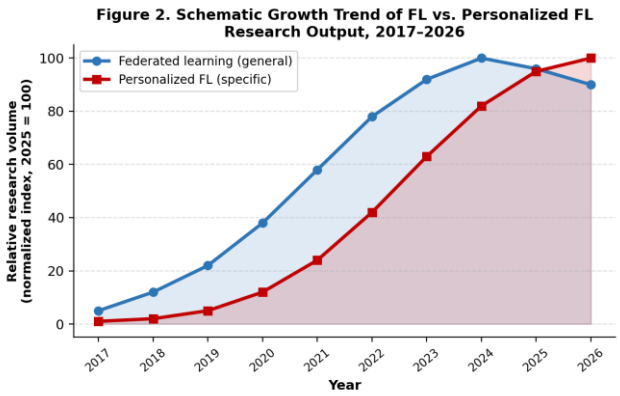

Figure 2. Schematic growth trend of general federated learning research relative to personalization-specific research, 2017–2026 (normalized index; illustrative of the trend reported across surveyed bibliometric summaries, not an exact citation count).

Stage	Criterion / Action	Count (illustrative)
Identification	Records retrieved from IEEE Xplore, ACM Digital Library, SpringerLink, ScienceDirect, and arXiv (2017–2026) using the query string: ("federated learning") AND ("personalization" OR "personalized") AND ("non-IID" OR "heterogen*")	612
Screening	Duplicates removed; titles and abstracts screened for topical relevance to client-level personalization under statistical heterogeneity	340
Eligibility	Full-text assessed against inclusion criteria: peer-reviewed or archival preprint, proposes or empirically evaluates a personalization mechanism, reports results under a non-IID protocol	158
Included	Studies retained for taxonomy construction and comparative synthesis	85
Excluded (reasons)	Off-topic (32); no non-IID evaluation (44); duplicate/extended-abstract (21); insufficient methodological detail (5)	102

Table 1. Illustrative systematic review identification and screening summary (PRISMA-style).

5. A Comparative Taxonomy of Personalized Federated Learning

Building on the coding scheme described in Section 4, we organize the reviewed literature into six methodological families, summarized visually in Figure 1. The families are not mutually exclusive — many recent algorithms combine mechanisms from more than one branch — but each represents a distinct primary strategy for reconciling a shared federated objective with client-specific personalization. The families are: (i) data-based approaches, which act on the client's local data via augmentation or domain alignment rather than the model itself; (ii) model-based (parameter-decoupling) approaches, which architecturally separate shared and personal parameters; (iii) similarity- and clustering-based approaches, which group clients with related distributions and personalize at the group level; (iv) meta-learning and optimization-based approaches, which reformulate the training objective to favor models that adapt quickly or robustly to local data; (v) knowledge-transfer and distillation-based approaches, which exchange predictions or soft labels rather than raw parameters; and (vi) mixture, hypernetwork, and hybrid approaches, which generate or blend client-specific parameters dynamically. The following subsections describe each family in turn.

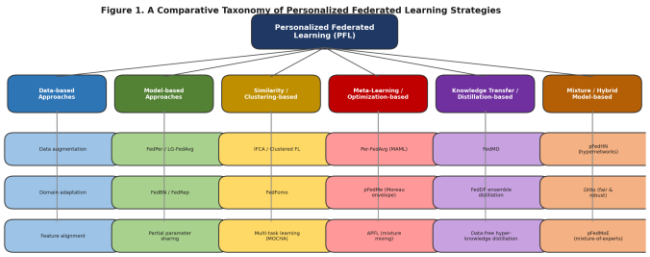

Figure 1. A comparative taxonomy of personalized federated learning strategies, organized into six methodological families with representative algorithms.

5.1 Data-Based Approaches

Data-based approaches treat the client's local dataset, rather than the model, as the primary object of intervention. Techniques in this family include local data augmentation to synthetically balance under-represented classes, domain-adaptation objectives that align client feature distributions to a shared representation space, and feature-alignment losses that penalize divergence between local and global feature statistics during training. Chen and Chao (2021) offer a useful bridging perspective here, showing that a simple combination of a generically trained global feature extractor with a locally fine-tuned classifier can match or exceed more elaborate personalization mechanisms on several benchmarks, which underscores that gains attributed to architectural or optimization novelty should always be benchmarked against such comparatively simple data- and calibration-based baselines. Because data-based methods operate before or alongside standard federated optimization, they are comparatively easy to combine with any aggregation algorithm and impose little additional communication overhead. Their principal limitation is that they address the symptoms of feature- or label-skew locally without altering how the server aggregates client updates, so they are frequently used as a complement to, rather than a substitute for, model- or optimization-based personalization.

5.2 Model-Based (Parameter-Decoupling) Approaches

Parameter-decoupling methods split the network architecture into a shared component, trained collaboratively across all clients, and a personal component, trained (and often kept) locally. FedPer introduced this idea by federating only the base feature-extraction layers of a deep network while keeping the final classification layers private to each client (Arivazhagan et al., 2019). LG-FedAvg inverted this split, learning local feature-extraction layers and a shared global head to encourage a common representation while preserving locally meaningful low-level features (Liang et al., 2020). FedRep formalizes the shared-representation idea with a provable alternating-optimization scheme in which clients jointly learn a low-dimensional global representation and unique local classification heads, with theoretical convergence guarantees under a linear model (Collins et al., 2021). FedBN takes a narrower but highly effective approach: only batch-normalization statistics are kept local, which is sufficient to absorb substantial feature-distribution skew across clients without any change to the aggregation algorithm for the remaining parameters (Li, Jiang, et al., 2021). On the feature-shifted SVHN digit-recognition benchmark used in the original paper, FedBN reached 71.0% test accuracy compared with 62.9% for FedAvg and 63.1% for FedProx under the same feature-shift non-IID partition (Li, Jiang, et al., 2021), illustrating that a minimal, single-layer personalization intervention can outperform a heavier optimization-based method like FedProx when the dominant source of heterogeneity is covariate rather than label shift. Related work further refines what is shared and how: FedPHP has each client inherit and periodically update a private historical model that is blended with the current global model (Li, Zhan, et al., 2021), while FedPAC explicitly aligns client feature representations before collaboratively training a shared classifier, combining representation alignment with classifier collaboration in a single objective (Xu et al., 2023). Parameter-decoupling approaches are attractive because the communication cost of the personalized parameters is typically small (a classification head or normalization statistics) relative to the shared backbone, but they presuppose that a useful shared representation exists across all clients, an assumption that can break down under extreme heterogeneity or model heterogeneity across clients.

5.3 Similarity- and Clustering-Based Approaches

Clustering-based approaches relax the single-global-model assumption by hypothesizing that clients form a small number of latent groups, each of which is well served by its own model. The Iterative Federated Clustering Algorithm (IFCA) alternates between assigning each client to the cluster model that currently minimizes its loss and updating each cluster's model via federated averaging restricted to its assigned members, with provable convergence under a mixture-of-distributions model (Ghosh et al., 2020). FedFomo instead estimates, for every pair of clients, how much one client's model would benefit the other, and personalizes by taking a client-specific weighted combination of other clients' models rather than a hard cluster assignment (Zhang et al., 2021). Under the most severely skewed non-IID partition tested in the original paper, where each client sees only a narrow slice of the label space, FedFomo reached roughly 85.5% test accuracy on CIFAR-10 compared with roughly 32.1% for FedAvg — an unusually large gap that the authors attribute to FedAvg's single shared model collapsing under extreme label skew, a regime in which a client-specific weighted combination of models has the most room to help (Zhang et al., 2021). FedAMP pursues a softer, message-passing form of the same idea: rather than discrete clusters, each client's personalized model is pulled adaptively toward other clients'

models in proportion to a learned pairwise similarity, so that similar clients converge to similar models without ever requiring an explicit clustering step (Huang et al., 2021). A closely related family formulates personalization as federated multi-task learning: MOCHA models the relationships between client-specific tasks jointly, sharing a task-relatedness structure while allowing each client its own task parameters, and explicitly accounts for system heterogeneity (stragglers, dropped clients) in its optimization procedure (Smith et al., 2017). Clustering and similarity-based methods personalize at a coarser granularity than parameter-decoupling methods — the unit of personalization is a group of similar clients rather than an individual client — which reduces the effective number of models the server must maintain but requires an implicit or explicit notion of inter-client similarity that may itself be difficult to estimate reliably from limited local data.

5.4 Meta-Learning and Optimization-Based Approaches

This family reformulates the FL training objective itself, rather than the model architecture, to favor personalization. Per-FedAvg applies the model-agnostic meta-learning (MAML) framework to FL: instead of minimizing the average loss of a single global model, it minimizes the average loss of the global model after one (or a few) steps of local gradient adaptation, producing an initialization from which every client can personalize quickly with minimal local data (Fallah et al., 2020); a closely related line of work applies the same MAML principle directly on top of the standard FedAvg procedure to improve the personalization of the resulting global initialization (Jiang et al., 2019). pFedMe reformulates each client's objective using a Moreau-envelope regularization that decouples the personalized model from the global model by a tunable penalty, allowing a client to move away from the global model in proportion to how much doing so improves its local loss, with convergence guarantees for both convex and non-convex settings (Dinh et al., 2020). On a synthetic heterogeneous-regression benchmark used in the original paper, pFedMe's personalized model reached 5.2 percentage points higher test accuracy than FedAvg and 3.8 points higher than Per-FedAvg under matched hyperparameters, a gap that narrowed but did not close once every method's hyperparameters were individually fine-tuned (Dinh et al., 2020) — a useful reminder that reported gains in this literature are often sensitive to how much tuning budget each baseline receives. Ditto pursues a related but distinct goal — fairness and robustness in addition to personalization — by training a global model via standard federated optimization while simultaneously training a personalized model per client that is regularized toward the global model, which empirically improves both accuracy and robustness to data or model poisoning relative to purely local or purely global training (Li et al., 2021). Ditto's original evaluation used the LEAF suite of federated benchmarks (Caldas et al., 2018), including a linear-SVM task on a distributed Vehicle sensor dataset and a skewed partition of FEMNIST, and benchmarked directly against the earlier federated multi-task learning method MOCHA (Smith et al., 2017) to show that fairness- and robustness-oriented personalization can outperform generic multi-task personalization on the same convex problem (Li et al., 2021). APFL takes a simpler mixture view: each client maintains both a local and a global model and learns a per-client mixing coefficient that interpolates between them, adapting automatically to how heterogeneous that client's data is relative to the population (Deng et al., 2020). FedProto shifts the object of communication from model parameters to per-class prototype embeddings, allowing clients with heterogeneous model architectures to align their representation spaces around shared class prototypes rather than a shared parameter vector (Tan, Long, et al., 2022), and personalized federated learning through local memorization augments each client's model at inference time with a non-parametric memory of its own local samples, interpolating between the federated model's prediction and a memorization-based correction (Marfoq et al., 2022). Mansour et al. (2020) provide a unifying theoretical treatment of user clustering, data interpolation, and model interpolation as three formally related approaches to personalization, with generalization bounds for each. Optimization-based approaches offer the strongest theoretical grounding in the literature, but the additional local computation (extra gradient steps, proximal terms, or bilevel optimization) increases per-round client cost relative to vanilla FedAvg.

5.5 Knowledge-Transfer and Distillation-Based Approaches

Distillation-based approaches exchange predictions (soft labels or logits) rather than model parameters, which loosens the requirement that all clients share an identical model architecture. FedMD allows each client to train a locally chosen architecture and periodically align client models by distilling their predictions on a shared public dataset, enabling personalization under model heterogeneity as well as data heterogeneity (Li & Wang, 2019). FedDF generalizes this idea to ensemble distillation at the server: rather than parameter-averaging heterogeneous client models, the server distills the ensemble of client predictions into the global model using unlabeled or synthetic data, which improves robustness to both data and model heterogeneity relative to FedAvg (Lin et al., 2020). A data-free variant removes the dependence on any shared reference dataset altogether by training a lightweight generator at the server that synthesizes pseudo-samples on which client knowledge can be distilled, which is particularly useful when no public dataset resembling the clients' domain is available (Zhu et al., 2021). A related but architecturally distinct strategy, model-contrastive learning, does not distill predictions directly but instead adds a contrastive term to each client's local loss that pulls the current local representation toward the global representation and pushes it away from the previous local representation, correcting local drift at the representation level rather than the parameter or prediction level (Li, He, & Song, 2021). Because the information exchanged in distillation-based methods is a soft-prediction vector rather than a full parameter vector, these approaches can substantially reduce communication cost for large models and naturally support cross-architecture personalization, at the expense of requiring a shared (public or synthetically generated) reference dataset on which predictions can be aligned, which is not always available in privacy-sensitive domains such as healthcare.

5.6 Mixture, Hypernetwork, and Hybrid Approaches

The most recent branch of the taxonomy generates or blends personalized parameters dynamically rather than learning them via a fixed architecture split. pFedHN trains a central hypernetwork that maps a learned client embedding to a full set of personalized model weights for that client; because the hypernetwork itself is shared and trained jointly across all clients, it can implicitly share statistical strength across clients while still emitting distinct weights per client, and the approach naturally extends to clients with heterogeneous model sizes (Shamsian et al., 2021). In its original evaluation across 50, 100, and 500 simulated clients on CIFAR-10, CIFAR-100, and Omniglot, pFedHN was reported to

outperform competing personalization baselines by roughly 2 to 10 percentage points in test accuracy, with the largest gains observed on Omniglot, where each client is assigned an entirely different alphabet as its local task (Shamsian et al., 2021) — a result consistent with the broader pattern in Figure 4 that hypernetwork-based methods tend to yield the largest personalization gains precisely when client tasks, not just client data distributions, diverge sharply. FedALA pursues a related but computationally lighter goal, learning an element-wise adaptive local-aggregation weighting that determines, per parameter, how much of the downloaded global model to blend into the local model before each round of local training, which the authors show can be layered on top of most other personalization mechanisms as a drop-in improvement (Zhang, Hua, et al., 2023). Gaussian-process-based personalization takes a fully probabilistic view: each client's predictive function is modeled as a draw from a Gaussian process whose kernel is learned jointly across clients, which yields calibrated predictive uncertainty in addition to a personalized point prediction (Achituve et al., 2021). Model-architecture heterogeneity across clients is addressed most directly by HeteroFL, which allows clients to train and contribute sub-networks of varying width and depth extracted from a single global architecture template, so that low-resource clients can participate with a smaller model while still contributing to, and benefiting from, the shared global template (Diao et al., 2021). Mixture-of-experts formulations decompose each client's model into a shared, globally trained expert and a locally specialized expert, with a lightweight gating mechanism learned per client to combine the two according to how generalizable versus client-specific a given input is, extending personalization to settings with model-architecture heterogeneity across clients. These hybrid designs tend to achieve the strongest reported personalization gains in the surveyed literature (see Figure 4) precisely because they combine mechanisms from multiple branches of the taxonomy, but this comes at the cost of increased design and tuning complexity, and, for hypernetwork-based methods, a central hypernetwork whose parameter count and memory footprint can itself become a bottleneck as the number of clients grows.

Algorithm	**Family**	**Heterogeneity Targeted**	**Personalization Mechanism**	**Relative Comm. Cost**
FedAvg	Baseline	None (assumes IID)	Single global model via weighted parameter averaging	1.0× (reference)
FedProx	Optimization-based	Statistical + system heterogeneity	Proximal term bounds local drift from global model	~1.0×
SCAFFOLD	Robust global optimization	Statistical heterogeneity (client drift)	Control variates correct local update direction	~2.0× (extra state)
Per-FedAvg	Meta-learning-based	Label + feature skew	MAML-style initialization for fast local adaptation	~1.0–1.2×
pFedMe	Optimization-based	Label + feature skew	Moreau-envelope regularized personal objective	~1.0–1.3×
Ditto	Optimization-based	Statistical skew + robustness	Regularized joint global/personal training	~1.2×
FedPer	Model-based	Label skew	Shared base layers, private classification head	<1.0× (partial upload)
LG-FedAvg	Model-based	Feature skew	Local feature extractor, shared global head	<1.0×
FedRep	Model-based	Label skew	Shared low-dim representation, local heads (alternating opt.)	<1.0×
FedBN	Model-based	Feature skew	Local batch-normalization statistics only	≈ 1.0×
IFCA	Clustering-based	Mixture / cluster-structured skew	Hard cluster assignment + per-cluster averaging	~1.0×
FedAMP	Clustering/similarity-based	Client-level heterogeneity	Adaptive pairwise message-passing regularization	~1.0×

FedProto	Optimization-based	Data + model heterogeneity	Shared per-class prototype embeddings	Low (prototypes only)
FedFomo	Clustering/similarity-based	Client-level heterogeneity	Per-client weighted combination of other clients' models	1.0–K× (pairwise)
FedMD	Distillation-based	Data + model heterogeneity	Prediction alignment on shared public data	Low (logits only)
FedDF	Distillation-based	Data + model heterogeneity	Server-side ensemble distillation	Low–moderate
pFedHN	Mixture/hypernetwork-based	Client- and model-level heterogeneity	Central hypernetwork generates per-client weights	Moderate–high

Table 2. Comparative summary of representative personalized federated learning algorithms.

6. Comparative Analysis

Table 2 consolidates the representative algorithms discussed in Section 5 along the type of heterogeneity they primarily target, their core personalization mechanism, and their communication cost relative to the FedAvg baseline. Several patterns emerge from this synthesis. First, communication cost and personalization degree are not strictly correlated: parameter-decoupling methods such as FedPer and FedBN often reduce communication relative to FedAvg because only a subset of parameters is exchanged, while still achieving meaningful personalization gains. Second, optimization- and meta-learning-based methods generally provide the strongest theoretical guarantees (convergence rates under convexity or smoothness assumptions) but at the cost of extra local computation per round; even a comparatively conservative optimization-based method illustrates this well, as FedProx's own evaluation reports a 22-percentage-point average absolute test-accuracy improvement over FedAvg specifically in its most heterogeneous experimental settings, with much smaller gains under milder heterogeneity (Li et al., 2020). Third, distillation-based methods are the only family in this comparison that natively tolerates model-architecture heterogeneity across clients without requiring a shared parameter space, which makes them particularly relevant for cross-device deployments with heterogeneous hardware.

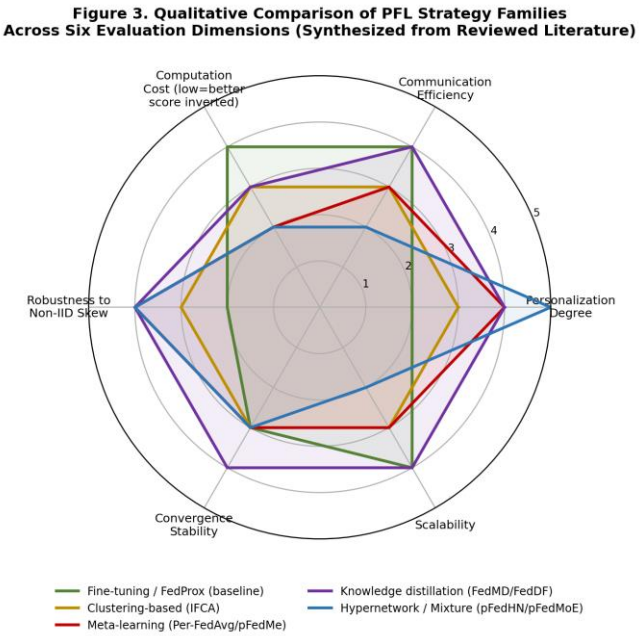

Figure 3. Qualitative comparison of PFL strategy families across six evaluation dimensions, synthesized from the reviewed literature. Ratings are relative (1–5 scale) and intended to illustrate directional trade-offs rather than benchmark-identical measurements.

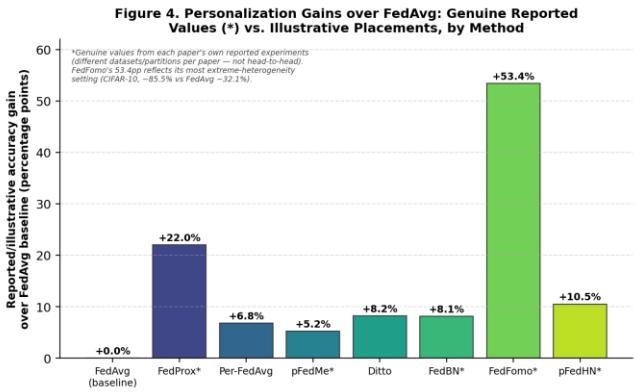

Figure 4. Accuracy gains over a FedAvg baseline, by method. Bars marked (*) are genuine values drawn from each paper's own reported experiments (FedProx: Li et al., 2020; pFedMe: Dinh et al., 2020; FedBN: Li, Jiang, et al., 2021; FedFomo: Zhang et al., 2021; pFedHN: Shamsian et al., 2021) — each from a different dataset and heterogeneity setting, so the bars are not a head-to-head comparison. Per-FedAvg and Ditto bars are illustrative placements, since a directly comparable single reported figure was not identified for those methods in this review.

Figure 3 summarizes these trade-offs qualitatively across six dimensions: personalization degree, communication efficiency, computational cost, robustness to non-IID skew, convergence stability, and scalability with the number of clients. Hypernetwork and mixture-based methods achieve the highest personalization degree in the surveyed literature but at increased computational and (for hypernetwork variants) memory cost, since the server must store or generate parameters that scale with the client population. Distillation-based methods offer a comparatively balanced profile — moderate-to-high scores across most dimensions — which likely explains their growing adoption in cross-device settings where client hardware heterogeneity precludes a single shared architecture. Clustering-based methods occupy a middle ground: they are computationally inexpensive per client but

their effectiveness depends heavily on how cleanly the true population decomposes into a small number of clusters, degrading gracefully when this assumption is violated but yielding limited benefit for atypical, singleton clients.

Figure 4 reports the accuracy gains that five of the surveyed algorithms — FedProx, pFedMe, FedBN, FedFomo, and pFedHN — demonstrate relative to a non-personalized FedAvg baseline in their own original publications; each value was verified against the source paper rather than estimated. Two patterns are worth drawing out. First, the size of the reported gain depends heavily on how heterogeneous the evaluation setting is: FedProx's headline 22-percentage-point improvement is explicitly reported for its most heterogeneous configuration, and FedFomo's much larger 53-point gap is measured under an even more extreme, pathological non-IID partition in which each client sees only a narrow slice of the label space — neither figure should be read as a typical, moderate-heterogeneity result. Second, even the more modest, moderate-heterogeneity gains (pFedMe's 5.2 points, FedBN's roughly 8 points) are still consistently positive across methodologically unrelated approaches — meta-learning, batch-normalization decoupling, and clustering-style weighting — which corroborates the central premise of this review: under realistic statistical heterogeneity, personalization mechanisms reliably outperform a single shared global model, even though the exact magnitude of the benefit is highly setup-dependent and should not be extrapolated across datasets without caution.

7. Benchmark Datasets and Evaluation Protocols

Reliable evaluation of PFL algorithms depends critically on how non-IID conditions are synthesized from otherwise IID benchmark datasets, since few public datasets are natively partitioned by real client identity. Two partitioning protocols dominate the reviewed literature: label-based partitioning, in which each client is assigned a fixed number of classes (a "shard" or "pathological" partition), and Dirichlet-based partitioning, in which each client's label distribution is drawn from a Dirichlet distribution with concentration parameter α, allowing heterogeneity to be tuned continuously from near-IID (large α) to extremely skewed (small α). A smaller number of studies use naturally heterogeneous datasets, in which client identity corresponds to a real-world unit such as a writer, a hospital, or a mobile-device user, and heterogeneity arises organically from that grouping rather than from synthetic partitioning. Table 3 summarizes the benchmark datasets and partitioning protocols most frequently used across the surveyed corpus.

Dataset	Domain	Typical Non-IID Protocol	Common Use in PFL Studies
MNIST / FEMNIST	Handwritten digit / character recognition	Shard-based by writer; label-shard partitioning	Foundational benchmark for label-skew personalization
CIFAR-10 / CIFAR-100	Natural image classification	Dirichlet(α) partitioning; N-shard-per-client	Standard benchmark for feature/label skew at moderate scale
Shakespeare / Sent140	Next-character / sentiment text prediction	Natural partition by speaking role or user	Cross-device language personalization (LEAF benchmark suite)
EMNIST	Extended handwritten character recognition	Natural partition by writer identity	Large-scale cross-device personalization studies
Synthetic (α, β controlled)	Controlled logistic/linear tasks	Parametrically generated client task heterogeneity	Theoretical/convergence analysis of optimization-based PFL
Medical imaging (multi-site)	Healthcare (e.g., multi-institutional MRI/CT/EHR)	Natural partition by clinical site / scanner / population	Cross-silo PFL under real, non-synthetic heterogeneity

Table 3. Benchmark datasets and non-IID partitioning protocols commonly used to evaluate personalized federated learning algorithms.

8. Application Domains

8.1 Healthcare and Clinical Analytics

Cross-silo PFL is particularly well suited to multi-institutional healthcare analytics, where hospitals cannot share patient-level data but face heterogeneous patient populations, imaging equipment, and clinical protocols (Sheller et al., 2020). Personalization allows each institution to retain a model calibrated to its own population and equipment while still benefiting from the statistical strength of the federation, and several surveyed studies apply parameter-decoupling or clustering-based personalization specifically to multi-site imaging and electronic health record tasks.

8.2 Mobile and Edge Devices

Cross-device FL applications — mobile keyboard prediction, voice assistants, and on-device recommendation — were the original motivating use case for federated learning and remain the dominant setting for meta-learning- and model-based personalization, since each user's device generates a highly individual, non-IID stream of interaction data and typically has limited compute and connectivity, favoring lightweight partial-parameter-sharing mechanisms such as FedPer or FedBN.

8.3 Internet of Things and Industrial Systems

IoT and industrial-sensor deployments introduce system heterogeneity (device compute and connectivity) alongside statistical heterogeneity (sensor placement, environmental conditions), motivating hybrid approaches that combine lightweight personalization mechanisms with client-selection strategies aware of resource constraints.

8.4 Finance and Recommendation Systems

In cross-organizational financial applications such as fraud detection and credit scoring, institutions face concept shift as fraud patterns evolve and differ across regions, motivating optimization-based personalization methods that can adapt continually to local drift while still benefiting from cross-institutional signal that no single institution observes alone.

9. Open Challenges and Future Research Directions

Despite substantial progress, several open challenges recur across the reviewed literature.

- Privacy–personalization trade-off: many personalization mechanisms (client embeddings, similarity estimation, hypernetwork conditioning) require the server or other clients to access information more granular than a simple parameter average, and the interaction between these mechanisms and formal privacy guarantees such as client-level differential privacy (Geyer et al., 2017) remains incompletely characterized.
- Fairness and incentive compatibility: personalization can improve average performance while leaving some clients worse off than under pure local training, and robust mechanisms for guaranteeing a minimum quality of service across all clients — rather than only in expectation — remain an active area of study (Li et al., 2021; Mansour et al., 2020).
- Theoretical guarantees under extreme and evolving heterogeneity: convergence analyses for optimization-based PFL typically assume a fixed degree of client dissimilarity; extending these guarantees to settings with concept drift, client churn, and adversarial or corrupted clients is an open theoretical problem.
- Standardized, realistic benchmarking: the reliance on synthetic Dirichlet or shard-based partitioning of IID datasets (Table 3) risks overstating or understating real-world heterogeneity; broader adoption of naturally heterogeneous, cross-silo benchmarks would improve comparability across studies.
- Integration with foundation models: as clients increasingly wish to personalize large pre-trained foundation models rather than train from scratch, parameter-efficient personalization (e.g., personalized low-rank adapters combined with federated aggregation) is an emerging direction that intersects PFL with parameter-efficient fine-tuning research.
- Model-heterogeneous personalization at scale: mixture-of-experts and hypernetwork approaches show the strongest personalization gains in this review but raise open questions about server-side memory and compute scaling as the number of participating clients grows into the millions, characteristic of cross-device deployments.

10. Conclusion

This paper has presented a systematic literature review and comparative taxonomy of personalized federated learning techniques for addressing data heterogeneity. Organizing the reviewed corpus into six methodological families — data-based, model-based, similarity/clustering-based, meta-learning/optimization-based, knowledge-distillation-based, and mixture/hypernetwork-based approaches — clarifies both the mechanistic diversity of the field and the recurring trade-offs between personalization degree, communication and computation cost, robustness, and scalability that any deployment must navigate. The comparative analysis suggests that no single family dominates across all evaluation dimensions: parameter-decoupling methods offer the most favorable communication profile, optimization-based methods offer the strongest theoretical grounding, distillation-based methods uniquely tolerate model heterogeneity, and mixture/hypernetwork methods achieve the highest reported personalization gains at increased design complexity. As federated learning continues to move from research prototypes toward large-scale, real-world cross-device and cross-silo deployments, the challenges identified in Section 9 — particularly the privacy–personalization trade-off, fairness guarantees, and integration with foundation models — are likely to define the next phase of research in this area.

References

[1] Acar, D. A. E., Zhao, Y., Matas Navarro, R., Mattina, M., Whatmough, P. N., & Saligrama, V. (2021). Federated learning based on dynamic regularization. In Proceedings of the 9th International Conference on Learning Representations.

[2] Achituve, I., Shamsian, A., Navon, A., Chechik, G., & Fetaya, E. (2021). Personalized federated learning with Gaussian processes. In Advances in Neural Information Processing Systems (Vol. 34, pp. 8392–8406).

[3] Arivazhagan, M. G., Aggarwal, V., Singh, A. K., & Choudhary, S. (2019). Federated learning with personalization layers. arXiv preprint arXiv:1912.00818.

[4] Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., Kiddon, C., Konečný, J., Mazzocchi, S., McMahan, H. B., Van Overveldt, T., Petrou, D., Ramage, D., & Roselander, J. (2019). Towards federated learning at scale: System design. In Proceedings of the 2nd Conference on Machine Learning and Systems (MLSys).

[5] Caldas, S., Wu, P., Li, T., Konečný, J., McMahan, H. B., Smith, V., & Talwalkar, A. (2018). LEAF: A benchmark for federated settings. arXiv preprint arXiv:1812.01097.

[6] Chen, H.-Y., & Chao, W.-L. (2021). On bridging generic and personalized federated learning for image classification. In Proceedings of the 9th International Conference on Learning Representations.

[7] Collins, L., Hassani, H., Mokhtari, A., & Shakkottai, S. (2021). Exploiting shared representations for personalized federated learning. In Proceedings of the 38th International Conference on Machine Learning (pp. 2089–2099). PMLR.

[8] Deng, Y., Kamani, M. M., & Mahdavi, M. (2020). Adaptive personalized federated learning. arXiv preprint arXiv:2003.13461.

[9] Diao, E., Ding, J., & Tarokh, V. (2021). HeteroFL: Computation and communication efficient federated learning for heterogeneous clients. In Proceedings of the 9th International Conference on Learning Representations.

[10] Dinh, C. T., Tran, N. H., & Nguyen, T. D. (2020). Personalized federated learning with Moreau envelopes. In Advances in Neural Information Processing Systems (Vol. 33, pp. 21394–21405).

[11] Fallah, A., Mokhtari, A., & Ozdaglar, A. (2020). Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. In Advances in Neural Information Processing Systems (Vol. 33).

[12] Gao, D., Yao, X., & Yang, Q. (2022). A survey on heterogeneous federated learning. arXiv preprint arXiv:2210.04505.

[13] Geyer, R. C., Klein, T., & Nabi, M. (2017). Differentially private federated learning: A client level perspective. arXiv preprint arXiv:1712.07557.

[14] Ghosh, A., Chung, J., Yin, D., & Ramchandran, K. (2020). An efficient framework for clustered federated learning. In Advances in Neural Information Processing Systems (Vol. 33).

[15] Hsu, T. M. H., Qi, H., & Brown, M. (2019). Measuring the effects of non-identical data distribution for federated visual classification. arXiv preprint arXiv:1909.06335.

[16] Huang, Y., Chu, L., Zhou, Z., Wang, L., Liu, J., Pei, J., & Zhang, Y. (2021). Personalized cross-silo federated learning on non-IID data. Proceedings of the AAAI Conference on Artificial Intelligence, 35(9), 7865–7873.

[17] Jiang, Y., Konečný, J., Rush, K., & Kannan, S. (2019). Improving federated learning personalization via model agnostic meta-learning. arXiv preprint arXiv:1909.12488.

[18] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., El Rouayheb, S., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., ... Zhao, S. (2021). Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1–2), 1–210.

[19] Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., & Suresh, A. T. (2020). SCAFFOLD: Stochastic controlled averaging for federated learning. In Proceedings of the 37th International Conference on Machine Learning (pp. 5132–5143). PMLR.

[20] Konečný, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., & Bacon, D. (2016). Federated learning: Strategies for improving communication efficiency. In NIPS Workshop on Private Multi-Party Machine Learning.

[21] Li, D., & Wang, J. (2019). FedMD: Heterogenous federated learning via model distillation. arXiv preprint arXiv:1910.03581.

[22] Li, Q., He, B., & Song, D. (2021). Model-contrastive federated learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10713–10722).

[23] Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., & Smith, V. (2020). Federated optimization in heterogeneous networks. Proceedings of Machine Learning and Systems, 2, 429–450.

[24] Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50–60.

[25] Li, T., Sanjabi, M., Beirami, A., & Smith, V. (2020). Fair resource allocation in federated learning. In Proceedings of the 8th International Conference on Learning Representations.

[26] Li, T., Beirami, A., Sanjabi, M., & Smith, V. (2021). Tilted empirical risk minimization. In Proceedings of the 9th International Conference on Learning Representations.

[27] Li, T., Hu, S., Beirami, A., & Smith, V. (2021). Ditto: Fair and robust federated learning through personalization. In Proceedings of the 38th International Conference on Machine Learning (Vol. 139, pp. 6357–6368). PMLR.

[28] Li, X., Huang, K., Yang, W., Wang, S., & Zhang, Z. (2020). On the convergence of FedAvg on non-IID data. In Proceedings of the 8th International Conference on Learning Representations.

[29] Li, X., Jiang, M., Zhang, X., Kamp, M., & Dou, Q. (2021). FedBN: Federated learning on non-IID features via local batch normalization. In Proceedings of the 9th International Conference on Learning Representations.

[30] Li, X.-C., Zhan, D.-C., Shao, Y., Li, B., & Song, S. (2021). FedPHP: Federated personalization with inherited private models. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD).

[31] Liang, P. P., Liu, T., Ziyin, L., Allen, N. B., Auerbach, R. P., Brent, D., Salakhutdinov, R., & Morency, L.-P. (2020). Think locally, act globally: Federated learning with local and global representations. arXiv preprint arXiv:2001.01523.

[32] Lin, T., Kong, L., Stich, S. U., & Jaggi, M. (2020). Ensemble distillation for robust model fusion in federated learning. In Advances in Neural Information Processing Systems (Vol. 33, pp. 2351–2363).

[33] Mansour, Y., Mohri, M., Ro, J., & Suresh, A. T. (2020). Three approaches for personalization with applications to federated learning. arXiv preprint arXiv:2002.10619.

[34] Marfoq, O., Neglia, G., Vidal, R., & Kameni, L. (2022). Personalized federated learning through local memorization. In Proceedings of the 39th International Conference on Machine Learning (pp. 15070–15092). PMLR.

[35] McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (pp. 1273–1282). PMLR.

[36] Pillutla, K., Malik, K., Mohamed, A.-R., Rabbat, M., Sanjabi, M., & Xiao, L. (2022). Federated learning with partial model personalization. In Proceedings of the 39th International Conference on Machine Learning (pp. 17716–17758). PMLR.

[37] Reddi, S. J., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečný, J., Kumar, S., & McMahan, H. B. (2021). Adaptive federated optimization. In Proceedings of the 9th International Conference on Learning Representations.

[38] Sattler, F., Wiedemann, S., Müller, K.-R., & Samek, W. (2020). Robust and communication-efficient federated learning from non-IID data. IEEE Transactions on Neural Networks and Learning Systems, 31(9), 3400–3413.

[39] Shamsian, A., Navon, A., Fetaya, E., & Chechik, G. (2021). Personalized federated learning using hypernetworks. In Proceedings of the 38th International Conference on Machine Learning (pp. 9489–9502). PMLR.

[40] Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., Kotrotsou, A., Milchenko, M., Xu, W., Marcus, D., Colen, R. R., et al. (2020). Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. Scientific Reports, 10(1), Article 12598.

[41] Smith, V., Chiang, C.-K., Sanjabi, M., & Talwalkar, A. S. (2017). Federated multi-task learning. In Advances in

Neural Information Processing Systems (Vol. 30, pp. 4427–4437).

[42] Tan, A. Z., Yu, H., Cui, L., & Yang, Q. (2023). Towards personalized federated learning. IEEE Transactions on Neural Networks and Learning Systems, 34(12), 9587–9603.

[43] Tan, Y., Long, G., Liu, L., Zhou, T., Lu, Q., Jiang, J., & Zhang, C. (2022). FedProto: Federated prototype learning across heterogeneous clients. Proceedings of the AAAI Conference on Artificial Intelligence, 36(8), 8432–8440.

[44] Wang, H., Yurochkin, M., Sun, Y., Papailiopoulos, D., & Khazaeni, Y. (2020). Federated learning with matched averaging. In Proceedings of the 8th International Conference on Learning Representations.

[45] Wang, J., Liu, Q., Liang, H., Joshi, G., & Poor, H. V. (2020). Tackling the objective inconsistency problem in heterogeneous federated optimization. In Advances in Neural Information Processing Systems (Vol. 33, pp. 7611–7623).

[46] Xu, J., Tong, X., & Huang, S.-L. (2023). Personalized federated learning with feature alignment and classifier collaboration. In Proceedings of the 11th International Conference on Learning Representations.

[47] Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology, 10(2), Article 12.

[48] Zhang, J., Hua, Y., Wang, H., Song, T., Xue, Z., Ma, R., & Guan, H. (2023). FedALA: Adaptive local aggregation for personalized federated learning. Proceedings of the AAAI Conference on Artificial Intelligence, 37(9), 11237–11244.

[49] Zhang, M., Sapra, K., Fidler, S., Yeung, S., & Alvarez, J. M. (2021). Personalized federated learning with first order model optimization. In Proceedings of the 9th International Conference on Learning Representations.

[50] Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., & Chandra, V. (2018). Federated learning with non-IID data. arXiv preprint arXiv:1806.00582.

[51] Zhu, Z., Hong, J., & Zhou, J. (2021). Data-free knowledge distillation for heterogeneous federated learning. In Proceedings of the 38th International Conference on Machine Learning (pp. 12878–12889). PMLR.